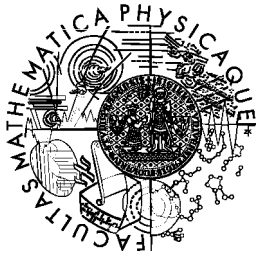


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Diana Kmeťková

Neparametrické testy nezávislosti

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	doc. RNDr. Zbyněk Pawlas, Ph.D.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2018

V prvom rade by som sa chcela poďakovať vedúcemu mojej bakalárskej práce doc. RNDr. Zbyňkovi Pawlasovi, Ph.D. za jeho čas, ochotu, cenné rady a pripomienky pri vzniku tejto bakalárskej práce. Ďalej by som sa tiež chcela poďakovať svojej rodine a blízkym za neustálu podporu pri štúdiu.

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Diana Kmeťková

Názov práce: Neparametrické testy nezávislosti

Autor: Diana Kmeťková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: doc. RNDr. Zbyněk Pawlas, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Cieľom tejto bakalárskej práce je prezentovať problém testovania nezávislosti dvoch náhodných veličín v neparametrickom modeli spojitých distribučných funkcií. Čitateľ sa najskôr oboznámi so základnými pojmami z teórie nezávislosti a z oblasti testov založených na poradí. Ďalej je predstavených pár najbežnejších metód na testovanie nezávislosti. Na začiatku je spomenutý jeden zástupca parametrických metód: test na základe Pearsonovho korelačného koeficientu, ďalej sa venujeme neparametrickým testom: testu založenom na Spearmanovom korelačnom koeficiente, Kendallovom korelačnom koeficiente a korelácii vzdialenosti. Podrobnejšie sme sa zamerali na Hoeffdingov test nezávislosti, ktorý je konzistentný voči všetkým alternatívam v modeli spojitých distribučných funkcií. Na záver sú pomocou simulácií v prostredí R porovnané jednotlivé štatistické metódy na testovanie nezávislosti.

Kľúčové slová: nezávislosť, neparametrické testy, testy založené na poradí, U-štatistiky, korelačný koeficient

Title: Nonparametric tests of independence

Author: Diana Kmeťková

Department: Department of Probability and Mathematical Statistics

Supervisor of the bachelor thesis: doc. RNDr. Zbyněk Pawlas, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The main objective of this thesis is the presentation regarding the problem of testing independence between two random variables in the nonparametric model of continuous cumulative distribution functions. Firstly, the reader is informed with basic notions from the theory of independence and rank tests. Afterwards, few of the most common methods for testing independence are introduced. In the beginning, the test based on Pearson's correlation coefficient is mentioned as a representative for parametric tests, then we continue with nonparametric tests, such as test based on Spearman's, Kendall's and distance correlation coefficient. We focus in better detail on Hoeffding's test of independence, which results to be consistent against all alternatives in the model of continuous cumulative distribution functions. In the end, we compare and evaluate presented methods for testing independence using simulations in R environment.

Keywords: independence, nonparametric tests, rank tests, U-statistics, correlation coefficient

Obsah

1 Úvod **2**

2 Teoretické základy **3**
2.1 Nezávislosť 3
2.2 Testy založené na poradiach 4
2.3 Trieda U-štatistík 5

3 Testovanie nezávislosti **7**
3.1 Úvod do problematiky testovania nezávislosti 7
3.1.1 Pearsonov korelačný koeficient 7
3.2 Neparametrické testy 8
3.2.1 Spearmanov korelačný koeficient 8
3.2.2 Kendallov korelačný koeficient 9
3.2.3 Korelácia vzdialenosti 10
3.3 Hoeffdingov test nezávislosti 11
3.3.1 Testová štatistika 13
3.3.2 Výpočet štatistiky D 14
3.3.3 Odvodenie rozptylu 14
3.3.4 D-test nezávislosti 18

4 Simulácie **19**
4.1 Postup 19
4.2 Výsledky 21
4.3 Diskusia 23

5 Záver **27**

Literatura **28**

Seznam obrázků **29**

Seznam tabulek **30**

1. Úvod

Nezávislosť je koncept, s ktorým sa v štatistike stretávame opakovane. Nezávislosť medzi veličinami môžeme interpretovať ako nedostatok určitého vzťahu či absenciu súvislosti. Inak povedané, zo znalosti hodnoty (výsledku pozorovania) jednej náhodnej veličiny nedokážeme usúdiť nič o druhej náhodnej veličine. Je to jedna z vlastností, ktoré nás zaujímajú v rámci štatistickej analýzy dát, keďže sa častokrát jedná o dôležitý predpoklad rôznych tvrdení či viet. Problém testovania nezávislosti je preto prítomný vo viacerých štúdiách a naprieč časom boli rozobrané rôzne metódy, ako testovať nezávislosť.

V bakalárskej práci sa zameriame na testovanie nezávislosti dvoch náhodných veličín a predstavíme najznámejšie metódy na testovanie ich závislosti, resp. nezávislosti. Hlavnou témou budú neparametrické testy, pričom konkrétnejšie opíšeme Hoeffdingov neparametrický test nezávislosti.

Na jednej strane máme v dvojrozmernom prípade klasické testy nezávislosti, medzi ktoré patrí napríklad test korelačného koeficientu. Za predpokladu normality sa jedná o najsilnejší nestranný a najsilnejší invariantný test nezávislosti v dvojrozmernom modeli. Ak sú potrebné podmienky splnené, parametrické testy sú spravidla silnejšie než neparametrické testy, avšak ich sila môže rapídne klesnúť, ak sa predpoklady porušia, ako ukázali Mudholkar a Wilding (2003).

Ponúka sa preto použitie neparametrických testov, medzi ktoré zaraďujeme test založený na Spearmanovom korelačnom koeficiente poradia, r', predstavený v Hotelling a Pabst (1936). Sami autori však upozorňujú, že problém merania rozsahu určitej závislosti či vzťahu medzi poradím, teda korelácie, sa líši od testovania existencie nejakého vzťahu, inak povedané absencie nezávislosti. Existenciu korelácie môžeme objaviť pomocou metódy založenej na poradí, ale tieto metódy samé o sebe nebudú nikdy postačujúce na testovanie miery závislosti.

Podobným problémom sa zaoberal aj Kendall (1938). Na meranie kompatibility dvoch poradí použil takzvané Kendallovo τ. Pre veľké hodnoty n sa rozdelenia r' a τ blížia k normálnemu, avšak rozdelenie τ je prekvapivo blízke normálnemu rozdeleniu aj pre malé hodnoty n. Nevýhodou týchto testov, ako preukázal Hoeffding (1948b), je, že nie sú asymptoticky nestranné vzhľadom k triede $\mathcal{F}$ distribučných funkcií so spojitými združenými a marginálnymi hustotami.

Testami, pri ktorých nepredpokladáme špeciálny tvar rozdelenia náhodného výberu, sa zaoberal aj Hoeffding (1948a), ktorý navrhol test nezávislosti založený na poradí, nazývaný aj D-test. Jedná sa o konzistentný test, a teda asymptoticky nestranný vzhľadom k triede spojitých distribučných funkcií, ktoré majú spojité združené a marginálne hustoty. Analyzovaním neparametrických testov pokračovali aj Blum a kol. (1961), ktorí nadviazali na výsledky Hoeffdingovej práce a vyriešili asymptotické rozdelenie testovej štatistiky D_n za nulovej hypotézy použitím asymptoticky ekvivalentnej, ale mierne odlišnej štatistiky označovanej B_n. Cena za menej predpokladov je zvyčajne menšia sila než spomínané parametrické testy.

V nasledujúcej kapitole sa oboznámime so základnými pojmami a tvrdeniami v rámci problematiky testovania nezávislosti, ďalej predstavíme metódy na testovanie nezávislosti a nakoniec pomocou simulácií dané testy medzi sebou porovnáme.

2. Teoretické základy

2.1 Nezávislosť

V bakalárskej práci sa budeme zaoberať problémom testovania nezávislosti dvoch náhodných veličín X a Y, definovaných na rovnakom pravdepodobnostnom priestore, s absolútne spojitými distribučnými funkciami, ktoré majú zároveň spojité združené a marginálne hustoty. Na začiatku definujeme pojem nezávislosti náhodných veličín a spomenieme základné vety, ktoré sa týkajú nezávislosti. Dôkazy k vetám v tejto podkapitole sú uvedené napríklad v Anděl (2007, str. 33–34).

Definícia 1 (nezávislosť náhodných veličín). *Náhodné veličiny $X_1, \ldots, X_n$ nazveme nezávislé, ak pre ľubovoľné borelovské množiny $B_1, \ldots, B_n$ platí vzťah*

$$P(\cap_{k=1}^{n}\{\omega : X_k(\omega) \in B_k\}) = \prod_{k=1}^{n} P(\{\omega : X_k(\omega) \in B_k\}).$$

Veta 1. *Nech $\mathbf{X} = (X_1, \ldots, X_n)$ má združenú distribučnú funkciu F a nech F_i je marginálna distribučná funkcia veličiny X_i, $i = 1, \ldots, n$. Potom $X_1, \ldots, X_n$ sú nezávislé práve vtedy, keď platí*

$$F(x_1, \ldots, x_n) = F_1(x_1) \cdots F_n(x_n) \text{ pre všetky } x_1, \ldots, x_n \in \mathbb{R}.$$

Veta 2. *Nech náhodné veličiny $X_1, \ldots, X_n$ majú združenú hustotu f a marginálne hustoty $f_1, \ldots, f_n$. Potom $X_1, \ldots, X_n$ sú nezávislé práve vtedy, keď platí*

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n) \text{ pre skoro všetky } x_1, \ldots, x_n \in \mathbb{R}.$$

Veta 3. *Nech náhodné veličiny $X_1, \ldots, X_n$ sú nezávislé náhodné veličiny s konečnými strednými hodnotami. Potom platí*

$$\mathbb{E}(X_1 \cdots X_n) = \mathbb{E}(X_1) \cdots \mathbb{E}(X_n).$$

Závislosť veličín X, Y sa častokrát meria pomocou korelačného koeficientu. Budeme ho využívať na overenie, či testy správne vyhodnotili určitú formu závislosti medzi náhodnými veličinami.

Definícia 2 (korelačný koeficient). *Nech X, Y sú náhodné veličiny, ktoré spĺňajú nerovnosti $0 < \mathrm{var}(X) < \infty$, $0 < \mathrm{var}(Y) < \infty$, potom korelačným koeficientom $\mathrm{corr}(X, Y)$ náhodných veličín X, Y rozumieme*

$$\rho_{X,Y} = \mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}.$$

Z definície je zrejmé, že $\rho_{X,Y} = \rho_{Y,X}$. Nech $a, b, c, d \in \mathbb{R}$ sú také čísla, pre ktoré platí $ac \neq 0$, potom

$$\rho_{aX+b,cY+d} = \begin{cases} \rho_{X,Y} & \text{pre } ac > 0, \\ -\rho_{X,Y} & \text{pre } ac < 0. \end{cases}$$

Táto vlastnosť nám hovorí, že korelačný koeficient sa pri lineárnej transformácii nezmení vôbec alebo sa zmení len znamienko. Základné vlastnosti korelačného koeficientu sú popísané v nasledujúcej vete.

Veta 4. *Pre korelačný koeficient platí* $-1 \leq \rho_{X,Y} \leq 1$. *Rovnosť* $\rho_{X,Y} = 1$ *nastáva práve vtedy, keď* $Y = a + bX$ *s pravdepodobnosťou 1, pričom* $b > 0$ *a rovnosť* $\rho_{X,Y} = -1$ *platí práve vtedy, keď* $Y = a + bX$ *s pravdepodobnosťou 1, kde* $b < 0$.

Lemma 5. *Nech* X, Y *sú nezávislé náhodné veličiny a* $\mathbb{E}\,X < \infty, \mathbb{E}\,Y < \infty$, *potom* $\operatorname{cov}(X, Y) = 0$.

Veličiny, ktorých kovariancia je nulová, sa nazývajú nekorelované. Ak sú náhodné veličiny nekorelované, nemusí to nutne znamenať, že sú aj nezávislé. Vlastnosť z lemmatu teda využijeme v prípade, keď budeme predpokladať nezávislosť náhodných veličín.

2.2 Testy založené na poradiach

Nech $X_1, \ldots, X_n$, $n \in \mathbb{N}$, je náhodný výber z jednorozmerného spojitého rozdelenia s distribučnou funkciou F a hustotou f vzhľadom k Lebesgueovej miere. Označme túto triedu distribučných funkcií $\mathcal{G}$. Keďže $X_1, \ldots, X_n$ sú nezávislé a majú spojité rozdelenie, potom

$$\mathsf{P}(X_i = X_j \text{ pre nejaké } i, j \in \{1, \ldots, n\}) = 0.$$

Definícia 3 (usporiadaný náhodný výber). *Zoradením všetkých náhodných veličín* $X_1, \ldots, X_n$ *od najmenšej po najväčšiu dostávame usporiadaný náhodný výber*

$$X_{(1)} < X_{(2)} < X_{(3)} < \cdots < X_{(n-1)} < X_{(n)}.$$

Symbolom $X_{(k)}$ *budeme rozumieť* k*-tu najmenšiu hodnotu medzi pozorovaniami* $X_1, \ldots, X_n$ *a nazývame ju* k*-ta poriadková štatistika.*

Definícia 4 (poradie). *Poradím náhodnej veličiny* X_i *rozumieme prirodzené číslo* $R_i \in \{1, \ldots, n\}$ *také, že* $X_i = X_{(R_i)}$.

Poznámka:

1. Symbolom $\mathcal{P}_n$, $n \in \mathbb{N}$, označíme množinu všetkých permutácií postupnosti $(1, \ldots, n)$.
2. $R_i = \sum_{j=1}^{n} \mathbf{1}_{\langle 0,\infty)}(X_i - X_j) = \sum_{j=1}^{n} \mathbf{1}\{X_i \geq X_j\}$.
3. Poriadkové štatistiky a poradia sú náhodné veličiny a tiež štatistiky.

Veta 6. *Náhodný vektor* $R = (R_1, \ldots, R_n)$ *naberá všetky hodnoty na množine* $\mathcal{P}_n$, *pričom každá z nich má pravdepodobnosť* $1/n!$.

Dôkaz. Lemma 13.1 v knihe van der Vaart (1998). ❑

Testy založené na poradí patria do skupiny neparametrických testov. Ich základnou myšlienkou je, že sa nepozeráme na konkrétne napozorované hodnoty, ale zaujíma nás výsledné poradie vo výbere. Testová štatistika je potom vypočítaná na základe poradia náhodných veličín.

Charakteristickou vlastnosťou týchto testov je, že ostávajú invariantné voči transformáciám, ktoré zachovávajú poradie veličín. Obecne, niektoré parametre rozdelenia F_X pôvodného náhodného výberu sa po transformáciách môžu zmeniť, napr. stredná hodnota, rozptyl či vyššie momenty. Pri odvodzovaní rozptylu Hoeffdingovej testovej štatistiky za nulovej hypotézy tiež využijeme transformáciu náhodného výberu. Budeme však vyžadovať, aby sa testová štatistika a určité charakteristiky pôvodného výberu nezmenili.

Ďalšou výhodou testov založených na poradí je, že v rámci neparametrických testov patria medzi jedny z najsilnejších. V prípade, že nie sú porušené predpoklady parametrických metód, sila testov založených na poradí je častokrát porovnateľná so silou parametrických testov. Na druhú stranu, ak predpoklady parametrických testov platiť nebudú, sila neparametrických testov zvykne byť väčšia, ako uviedol Hoeffding (1948a).

2.3 Trieda U-štatistík

Vďaka svojím vlastnostiam sú U-štatistiky základom mnohých testových štatistík v oblasti parametrickej a neparametrickej štatistiky. V tejto časti definujeme pojem U-štatistiky a spomenieme niektoré z vlastností, ktoré budeme používať pri odvodení Hoeffdingovho testu nezávislosti v kapitole 3.

Predpokladajme, že máme náhodný výber $X_1, \ldots, X_n$ charakterizovaný distribučnou funkciou $F \in \mathcal{G}$. Uvažujme funkcionál $\theta(F)$ s definičným oborom $\mathcal{G}$ a hodnotami v $\mathbb{R}$.

Definícia 5 (regulárny funkcionál a jadro). *Povieme, že $\theta = \theta(F)$ je regulárny funkcionál na $\mathcal{G}$, ak pre všetky distribučné funkcie $F \in \mathcal{G}$ existuje nestranný odhad $\theta(F)$. V tomto prípade platí, že pre všetky $F \in \mathcal{G}$ môžeme $\theta(F)$ zapísať nasledovne:*

$$\theta(F) = \mathbb{E}(\Phi(X_1, \ldots, X_m)) = \int \cdots \int \Phi(x_1, \ldots, x_m) dF(x_1) \cdots dF(x_m)$$

pre nejakú funkciu $\Phi = \Phi(x_1, \ldots, x_m)$. Takúto funkciu budeme nazývať jadro funkcionálu $\theta(F)$.

Definícia 6 (U-štatistika). *Nech $n \geq m$. Pre ľubovoľné jadro $\Phi(x_1, \ldots, x_m)$ definujeme príslušnú U-štatistiku pre odhad $\theta(F)$ ako priemer jadra $\Phi(x_1, \ldots, x_m)$ cez všetky permutácie pozorovaní:*

$$U_n = U_n(\Phi) = \frac{(n-m)!}{n!} \sum_{P_{m,n}} \Phi(X_{i_1}, \ldots, X_{i_m}) =$$

$$= \frac{1}{n(n-1)\cdots(n-m+1)} \sum_{P_{m,n}} \Phi(X_{i_1}, \ldots, X_{i_m}),$$

kde súčet prechádza cez všetkých $\frac{n!}{(n-m)!}$ permutácií $(i_1, \ldots, i_m)$ veľkosti m z prvkov množiny $\{1, \ldots, n\}$. Ak je naviac jadro invariantné vzhľadom k permutáciám svojich argumentov, inak povedané, ak je Φ symetrické, potom U_n môžeme vyjadriť

$$U_n = U_n(\Phi) = \binom{n}{m}^{-1} \sum_{R_{m,n}} \Phi(X_{i_1}, \ldots, X_{i_m}), \tag{2.1}$$

kde súčet prechádza cez všetkých $\binom{n}{m}$ *usporiadaných m-tic* $i_1 < \cdots < i_m$ *vybraných z množiny* $\{1, \ldots, n\}$.

Nasledujúce lemma a poznámky sú viac rozobrané a dokázané v Hoeffding (1948b, str. 293–297) a Halmos (1946, str. 34–43).

Lemma 7. *Ak je* $\Phi(x_1, \ldots, x_m)$ *jadrom regulárneho funkcionálu* $\theta(F)$ *definovanom na* $\mathcal{G}$*, potom je štatistika* U_n *nestranným odhadom* $\theta(F)$ *na* $\mathcal{G}$*:*

$$\theta(F) = \int \cdots \int U_n(x_1, \ldots, x_n) dF(x_1) \cdots dF(x_n) = \mathbb{E}U_n$$

pre každú F *z* $\mathcal{G}$

Ľahko môžeme nahliadnuť, že každé nesymetrické jadro sa dá nahradiť symetrickou verziou Φ'.

Poznámka:
Nech Φ je jadro. Potom existuje symetrická štatistika $\Phi'(X_1, \ldots, X_m)$ definovaná nasledovne

$$\Phi'(X_1, \ldots, X_m) = \frac{1}{m!} \sum_{P_m} \Phi(X_{i_1}, \ldots, X_{i_m}), \tag{2.2}$$

kde suma prechádza cez všetkých $m!$ permutácií $i_1 < \cdots < i_m$ vybraných z množiny $\{1, \ldots, m\}$.

Jedná sa o priemer $m!$ foriem, z ktorých každá je nestranným odhadom parametru. Vďaka vlastnostiam strednej hodnoty máme, že symetrická funkcia $\Phi'(X_1, \ldots, X_m)$ je tiež nestranným odhadom, a teda aj jadrom parametru $\theta(F)$.

Poznámka:
Pre $n = m$, U_n sa skráti na symetrické jadro funkcionálu $\theta(F)$ z (2.2).

3. Testovanie nezávislosti

3.1 Úvod do problematiky testovania nezávislosti

V tejto sekcii stručne opíšeme niektoré z často používaných metód na testovanie závislosti a nezávislosti medzi náhodnými veličinami, ktoré sme vybrali na simuláčnú štúdiu. Nech $(X_1, Y_1), \ldots, (X_n, Y_n)$, $n \in \mathbb{N}$, je náhodný výber z rozdelenia z triedy $\mathcal{F}$, potom uvažujme nulovú hypotézu H_0: X a Y sú nezávislé, proti alternatíve H_1: neplatí H_0.

Začneme parametrickým testom, ktorý je založený na Pearsonovom korelačnom koeficiente a ktorý sme vybrali, aby sme mali aspoň jedného zástupcu parametrickej metódy na porovnanie výkonnosti s neparametrickými testami. Zvyšné testy patria do skupiny neparametrických testov, pričom sa zameriame na neparametrický test nezávislosti navrhnutý v Hoeffding (1948a).

3.1.1 Pearsonov korelačný koeficient

Jeden z hlavných parametrov, ktorý sa pri testovaní určitého vzťahu medzi veličinami používa, je korelačný koeficient (definícia 2). Jeho konzistentným odhadom je výberový korelačný koeficient, nazývaný tiež Pearsonov.

Definícia 7. *Nech* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ *a* $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ *sú výberové priemery náhodného výberu* $\mathbf{X} = (X_1, \ldots, X_n)$, *resp.* $\mathbf{Y} = (Y_1, \ldots, Y_n)$. *Potom Pearsonov korelačný koeficient je definovaný nasledovne:*

$$\rho_{Pearson}(\mathbf{X}, \mathbf{Y}) = \rho_p = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}.$$

V nasledujúcom tvrdení ukážeme rozdelenie Pearsonovho korelačného koeficientu za predpokladu normality.

Tvrdenie 8. *Ak máme náhodný výber z dvojrozmerného normálneho rozdelenia, ktorý má kladné rozptyly a korelačný koeficient* $\rho = 0$, *potom pre* $n \geq 3$ *platí, že*

$$T_n = \sqrt{n-2}\frac{\rho_p}{\sqrt{1-\rho_p^2}} \sim t_{n-2}.$$

Dôkaz. Veta 6.2 v Anděl (2007). ❑

Ak budeme predpokladať nezávislosť medzi $\mathbf{X}$ a $\mathbf{Y}$, potom podľa lemmatu 5 platí, že $\rho = 0$ a dá sa ukázať, že

$$T_n = \sqrt{n-2}\frac{\rho_p}{\sqrt{1-\rho_p^2}} \xrightarrow[n\to\infty]{d} N(0,1).$$

Preto testovú štatistiku T_n môžeme použiť pre asymptotický test hypotézy H_0 proti H_1 aj v prípade, že náhodný výber nepochádza z dvojrozmerného normálneho rozdelenia. Dostaneme test, ktorý bude citlivý proti alternatívam, pre ktoré

je skutočný korelačný koeficient ρ nenulový, takže medzi veličinami pozorujeme lineárnu závislosť. Na druhú stranu, ak X_i a Y_i nie sú nezávislé, ale sú nekorelované, takže $\rho = 0$, tento test konzistentný nebude, čo môže nastať napríklad, keď $Y_i = X_i^2$.

3.2 Neparametrické testy

Neparametrické testy sa používajú na testovanie hypotéz, keď nepoznáme presné rozdelenie dát. Pri takýchto testoch nie je nutné klásť predpoklady na funkčnú formu rozdelenia danej populácie. Ak nebude povedané inak, pri nasledujúcich testoch budeme predpokladať, že $(X_1, Y_1), \ldots, (X_n, Y_n)$ je náhodný výber zo spojitého dvojrozmerného rozdelenia.

3.2.1 Spearmanov korelačný koeficient

Obecne, hodnotenie výberového korelačného koeficientu je spojené so splnením predpokladu, že výber pochádza z normálneho rozdelenia. V praxi sa častokrát stretávame s tým, že tento predpoklad je porušený. Ďalším problémom môže byť, že nie sme schopní presne určiť hodnoty náhodných veličín a sú nám známe len ich poradia.

Definícia 8. *Nech $R_1, \ldots, R_n$ sú poradia veličín $X_1, \ldots, X_n$ a $S_1, \ldots, S_n$ sú poradia veličín $Y_1, \ldots, Y_n$. Potom Spearmanov korelačný koeficient je definovaný nasledovne:*

$$\rho_{Spearman}(R, S) = \rho_s = \frac{\sum_{i=1}^{n}(R_i S_i - n\bar{R}_n\bar{S}_n)}{\sqrt{(\sum_{i=1}^{n} R_i^2 - n\bar{R}_n^2)(\sum_{i=1}^{n} S_i^2 - n\bar{S}_n^2)}},$$

kde $\bar{R}_n$ a $\bar{S}_n$ sú výberové priemery poradí $R_1, \ldots, R_n$ a $S_1, \ldots, S_n$.

Výber $(X_1, Y_1), \ldots, (X_n, Y_n)$ pochádza zo spojitého rozdelenia, preto sa predpokladá, že v dátach nebudú zhody, tj. všetky hodnoty $X_1, \ldots, X_n$ budú navzájom odlišné a tak isto všetky hodnoty $Y_1, \ldots, Y_n$ budú odlišné. Potom môžeme písať

$$\bar{R}_n = \bar{S}_n = \frac{n+1}{2}, \tag{3.1}$$

$$\sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} S_i^2 = \frac{n(n+1)(2n+1)}{6}. \tag{3.2}$$

Treba podotknúť, že Spearmanov korelačný koeficient nie je odhadom teoretického korelačného koeficientu z definície 2, ale odhaduje

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(F_X(X_i), F_Y(Y_i))}{\sqrt{\mathrm{var}(F_X(X_i))}\sqrt{\mathrm{var}(F_Y(Y_i))}},$$

kde F_X a F_Y sú distribučné funkcie náhodných veličín X_i a Y_i.

Použitím (3.1) a (3.2) vieme Spearmanov korelačný koeficient vyjadriť v tvare

$$\rho_s = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - S_i)^2.$$

Ak je X_i nezávislé s Y_i pre každé i, potom by sa táto nezávislosť mala preniesť aj na ich poradia R_i a S_i. Taktiež si všimneme, že Spearmanov korelačný koeficient dosahuje maximálnu hodnotu 1 práve vtedy, keď $R_i = S_i$ pre každé i. To nastáva v situácii, keď existuje ostro rastúca funkcia g taká, že $X_i = g(Y_i)$ pre každé i, inak povedané, keď je X_i ostro rastúcou transformáciou Y_i. Na druhú stranu, tento koeficient naberá hodnotu -1 práve vtedy, keď existuje ostro klesajúca funkcia g taká, že $X_i = g(Y_i)$ pre každé i. V porovnaní s Pearsonovým korelačným koeficientom, ktorý nadobúda hodnotu 1 len v prípade, keď X_i je ostro rastúcou lineárnou transformáciou Y_i, Spearmanov koeficient nevyžaduje predpoklad linearity, a teda je schopný rozoznať aj iné typy závislosti. Ako je ukázané v Anděl (2007, str. 256–257) za H_0 máme

$$\mathbb{E}\,\rho_s = 0 \text{ a } \operatorname{var}\rho_s = \frac{1}{n-1}$$

a zároveň

$$\sqrt{n-1}\rho_s \xrightarrow[n\to\infty]{d} N(0,1).$$

Spearmanov korelačný koeficient môžeme využiť ku konštrukcii testu nezávislosti H_0, kde hypotézu zamietame, ak $\sqrt{n-1}|\rho_s| \geq u_{1-\alpha/2}$, kde $u_{1-\alpha/2}$ je $1-\alpha/2$ kvantil normovaného normálneho rozdelenia. Jedná sa o asymptotický test, ktorý nepredpokladá normalitu rozdelenia. Test je citlivý voči alternatívam, keď je korelačný koeficient medzi $F_X(X_i)$ a $F_Y(Y_i)$ nenulový, ale nie je konzistentný v prípadoch, keď X_i a Y_i nie sú nezávislé, ale korelačný koeficient je rovný 0. Tieto poznatky sú podrobnejšie rozobrané v Omelka (2018, str. 137–138).

3.2.2 Kendallov korelačný koeficient

Metóda testovania pomocou Kendallovho korelačného koeficientu je podobná Spearmanovej metóde, pretože je taktiež založená na poradí a dokáže identifikovať monotónne transformácie.

Majme náhodný výber $(X_1, Y_1), \ldots, (X_n, Y_n)$. Potom môžeme dvojice pozorovaní (x_i, y_i) a (x_j, y_j), kde $i \neq j$ nazvať súhlasné, ak sa poradia oboch prvkov zhodujú, tj. ak platí buď $x_i > x_j$ a zároveň $y_i > y_j$ alebo $x_i < x_j$ a zároveň $y_i < y_j$. V opačnom prípade, ak $x_i > x_j$ a zároveň $y_i < y_j$ alebo $x_i < x_j$ a zároveň $y_i > y_j$, takéto dvojice nazveme nesúhlasné. V prípade, že $x_i = x_j$ a $y_i = y_j$, takúto dvojicu neoznačíme za súhlasnú ani nesúhlasnú.

Definícia 9. *Kendallov korelačný koeficient, nazývaný tiež Kendallovo τ je definovaný nasledovne:*

$$\rho_{Kendall} = \tau = \frac{\textit{(počet súhlasných párov)} - \textit{(počet nesúhlasných párov)}}{n(n-1)/2}.$$

Ekvivalentne môžeme τ vyjadriť aj

$$\tau = \frac{1}{n(n-1)} \sum_{i \neq j} \operatorname{sgn}(X_i - X_j) \operatorname{sgn}(Y_i - Y_j)$$

Z definície je zrejmé, že $-1 \leq \tau \leq 1$. Kendallovo τ dosahuje hodnotu 1, keď poradia oboch výberov sú zhodné. Interpretácia je teda rovnaká ako pri Spearmanovom korelačnom koeficiente.

Ako ukázal Kendall (1938) hypotézu H_0 môžeme testovať transformovaním τ na testovaciu štatistiku $Z_\tau = \frac{\tau}{\sigma_\tau}$, kde $\sigma_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$ je smerodatná odchýlka τ. Táto štatistika má potom asymptoticky normálne normované rozdelenie.

3.2.3 Korelácia vzdialenosti

Metóda založená na korelácii vzdialenosti *dCor* bola predstavená v Székely a kol. (2007). V porovnaní s predchádzajúcimi metódami sa jedná o novodobejší prístup k testovaniu nezávislosti. Stále sa však vyznačuje vlastnosťami podobnými tým, ktoré by sme čakali od korelačných koeficientov.

1. Platí $0 \leq dCor \leq 1$.
2. $dCor = 0$ práve vtedy, keď sú X a Y nezávislé.

V prípade náhodného výberu z dvojrozmerného normálneho rozdelenia *dCor* je funkciou ρ a platí $dCor \leq |\rho(X, Y)|$, pričom rovnosť nastáva, keď $\rho = \pm 1$.

Definícia 10. *Nech* $\mathbf{X}$ *a* $\mathbf{Y}$ *sú dva náhodné vektory, ktoré majú konečné prvé momenty,* $f_{\mathbf{X},\mathbf{Y}}$ *je združená hustota a* $f_{\mathbf{X}}$, $f_{\mathbf{Y}}$ *sú marginálne hustoty. Nech* $\mathbf{X}_i \in \mathbb{R}^p$, $\mathbf{Y}_i \in \mathbb{R}^q$, $p, q \in \mathbb{N}$, *pre* $i \in \{1, \ldots, n\}$. *Vzdialenostná kovariancia, dCov, medzi* $\mathbf{X}$ *a* $\mathbf{Y}$ *je nezáporné číslo* $V(\mathbf{X}, \mathbf{Y})$ *také, že*

$$\begin{aligned} V^2(\mathbf{X}, \mathbf{Y}) &= \|f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) - f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{y})\|^2 \\ &= \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) - f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}}(\mathbf{y})|^2}{|\mathbf{x}|^{1+p} |\mathbf{y}|^{1+q}} \, d\mathbf{x} \, d\mathbf{y}, \end{aligned}$$

kde

$$c_p = \frac{\pi^{(1+p)/2}}{\Gamma((1+p)/2)}, \quad c_q = \frac{\pi^{(1+q)/2}}{\Gamma((1+q)/2)}.$$

Obdobne definujeme vzdialenostný rozptyl, dVar, ako odmocninu z výrazu

$$V^2(\mathbf{X}) = V^2(\mathbf{X}, \mathbf{X}) = \|f_{\mathbf{X},\mathbf{X}}(\mathbf{x}, \mathbf{y}) - f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{y})\|^2.$$

Definícia 11. *Vzdialenostná korelácia* [1]*, dCor, medzi dvomi náhodnými vektormi* $\mathbf{X}$ *a* $\mathbf{Y}$*, ktoré majú konečné prvé momenty, je nezáporné číslo* $R(\mathbf{X}, \mathbf{Y})$ *definované nasledovne*

$$R^2(\mathbf{X}, \mathbf{Y}) = \begin{cases} \frac{V^2(\mathbf{X},\mathbf{Y})}{\sqrt{V^2(\mathbf{X}) V^2(\mathbf{Y})}}, & \text{ak } V^2(\mathbf{X}) V^2(\mathbf{Y}) > 0, \\ 0, & \text{ak } V^2(\mathbf{X}) V^2(\mathbf{Y}) = 0. \end{cases}$$

[1] z anglického distance correlation

Nech $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ je náhodný výber z rozdelenia vektoru $(\mathbf{X}, \mathbf{Y})$. Potom definujeme [2]

$$
\begin{aligned}
a_{kl} &= |\mathbf{X}_k - \mathbf{X}_l|_p, & b_{kl} &= |\mathbf{Y}_k - \mathbf{Y}_l|_q \\
\bar{a}_{k.} &= \frac{1}{n}\sum_{l=1}^{n} a_{kl}, & \bar{a}_{.l} &= \frac{1}{n}\sum_{k=1}^{n} a_{kl}, \\
\bar{a}_{..} &= \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}, & A_{kl} &= a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..},
\end{aligned}
$$

kde $k, l = 1, \ldots, n$.

Obdobne sa definuje $\bar{b}_{k.}$, $\bar{b}_{.l}$, $\bar{b}_{..}$, potom $B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}_{..}$. Empirická vzdialenostná kovariancia $V_n(\mathbf{X}, \mathbf{Y})$ bude potom splňovať

$$V_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}.$$

Podobne

$$V_n^2(\mathbf{X}) = \sum_{k,l=1}^{n} A_{kl}^2 \quad \text{a} \quad V_n^2(\mathbf{Y}) = \sum_{k,l=1}^{n} B_{kl}^2.$$

Testová štatistika bude mať teda tvar

$$R_n^2(\mathbf{X}, \mathbf{Y}) = \begin{cases} \frac{V_n^2(\mathbf{X},\mathbf{Y})}{\sqrt{V_n^2(\mathbf{X})V_n^2(\mathbf{Y})}}, & \text{ak } V_n^2(\mathbf{X})V_n^2(\mathbf{Y}) > 0, \\ 0, & \text{ak } V_n^2(\mathbf{X})V_n^2(\mathbf{Y}) = 0. \end{cases}$$

Asymptotické rozdelenie testovej štatistiky je odvodené v Székely a kol. (2007).

3.3 Hoeffdingov test nezávislosti

V tejto práci sa zameriavame na problém testovania nezávislosti dvoch náhodných veličín X, Y, pričom nás bude zaujímať výkon Hoeffdingovho testu v porovnaní s vyššie spomínanými testami.

Nech $n \in \mathbb{N}$ a $(X_1, Y_1), \ldots, (X_n, Y_n)$ je náhodný výber z dvojrozmerného spojitého rozdelenia s distribučnou funkciou $F_{X,Y}(x, y)$ a hustotou $f_{X,Y}(x, y)$ vzhľadom k Lebesgueovej miere. Označme $\mathcal{F}$ triedu všetkých spojitých distribučných funkcií $F_{X,Y}(x, y)$, ktoré majú spojité združené a marginálne hustoty $f_{X,Y}(x, y)$, $f_X(x)$ a $f_Y(y)$.

Nulová hypotéza bude mať tvar

$$H_0 : F_{X,Y}(x, y) = F_X(x)F_Y(y) \text{ pre každé } x, y \in \mathbb{R} \tag{3.3}$$

proti alternatíve

$$H_1 : F_{X,Y}(x, y) \neq F_X(x)F_Y(y) \text{ pre nejaké } x, y \in \mathbb{R}.$$

Chceme testovať nulovú hypotézu (3.3) tak, aby $F \in \omega_0$, kde $\omega_0 \subset \mathcal{F}$ značí všetky rozdelenia v modeli $\mathcal{F}$, ktoré splňujú hypotézu H_0. Ak bude existovať

[2] $|\cdot|$ značí Euklidovskú normu

1. funkcionál $\theta(F)$ definovaný pre každú F z $\mathcal{F}$ taký, že $\theta(F) = 0$ práve vtedy, keď $F \in \omega_0$ a

2. konzistentný odhad $\theta(F)$,

potom ako ukázal Hoeffding (1948a) môžeme zostrojiť konzistentný test H_0.

Označme [3]

$$D(x, y) = F_{X,Y}(x, y) - F_X(x)F_Y(y)$$

a

$$\theta = \theta(F) = \int D^2(x, y)dF_{X,Y}(x, y) = \mathbb{E}[D(X, Y)]^2 =$$
$$= \text{var}(D(X, Y)) + (\mathbb{E}[D(X, Y)])^2.$$

Je zrejmé, že náhodné veličiny X, Y sú nezávislé práve vtedy, keď $D(x, y) = 0$ pre všetky $x, y \in \mathbb{R}$.

Veta 9. *Nech $F_{X,Y}(x, y) \in \mathcal{F}$. Potom $\theta(F) = 0$ práve vtedy, keď $D(x, y) \equiv 0$. (Ide o Theorem 3.1 z Hoeffding (1948a).)*

Dôkaz. Očividne, ak $D(x, y) \equiv 0$, potom $\theta(F) = 0$.

Predpokladajme, že $D(x, y) \not\equiv 0$. Označme $d(x, y) = f_{X,Y}(x, y) - f_X(x)f_Y(y)$. Keďže $F_{X,Y}(x, y) \in \mathcal{F}$, potom funkcia $d(x, y)$ je spojitá. Ďalej

$$D(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} d(u, v)dudv.$$

Predpoklad $D(x, y) \not\equiv 0$ implikuje, že $d(x, y) \not\equiv 0$

Z vlastnosti hustoty dostávame

$$\int\int d(x, y)dxdy = \int\int (f_{X,Y}(x, y) - f_X(x)f_Y(y))dxdy = 0,$$

a teda existuje nejaké okolie V v $\mathbb{R}^2$ také, že $d(x, y) > 0$ na V. Potom $D(x, y) \neq 0$ skoro všade na V a

$$d(x, y) > 0 \Leftrightarrow f_{X,Y}(x, y) - f_X(x)f_Y(y) > 0 \Leftrightarrow f_{X,Y}(x, y) > f_X(x)f_Y(y) \geq 0,$$

preto $f_{X,Y}(x, y) > 0$ na V. Dostávame

$$\theta(F) \geq \int\int_V D^2(x, y)f_{X,Y}(x, y)dxdy > 0,$$

čo je spor.

❑

[3] Ak nebude označený obor integrácie, máme na mysli integráciu cez celý priestor. V tomto prípade $\mathbb{R}^2$.

3.3.1 Testová štatistika

Na úvod zavedieme nasledovné funkcie. Nech

$$C(u) = \begin{cases} 1, & \text{ak } u \geq 0, \\ 0, & \text{ak } u < 0, \end{cases} \tag{3.4}$$

$$\psi(x_1, x_2, x_3) = C(x_1 - x_2) - C(x_1 - x_3),$$

$$\phi(x_1, y_1; \ldots; x_5, y_5) = \frac{1}{4}\psi(x_1, x_2, x_3)\psi(x_1, x_4, x_5)\psi(x_1, x_2, x_3)\psi(x_1, x_2, x_3).$$

Potom

$$\begin{aligned} \Delta &= \int \cdots \int \phi(x_1, y_1; \ldots; x_5, y_5) dF(x_1, y_1) \cdots dF(x_5, y_5) \\ &= \mathbb{E}[\phi(X_1, Y_1; \ldots; X_5, Y_5)]. \end{aligned}$$

Pre všetky $F \in \mathcal{F}$ existuje nestranný odhad Δ, takže z definície 5 je Δ regulárnym funkcionálom nad $\mathcal{F}$ a $\phi(x_1, y_1; \ldots; x_5, y_5)$ je jeho jadro.

Nech $n \geq 5$, potom definujeme testovú štatistiku

$$D_n = \frac{1}{n(n-1)\cdots(n-4)} \sum_{P_{5,n}} \phi(X_{\alpha_1}, Y_{\alpha_1}; \ldots; X_{\alpha_5}, Y_{\alpha_5}), \tag{3.5}$$

kde súčet prechádza cez všetky α_i také, že

$$\alpha_i = 1, \ldots, n; \qquad \alpha_i \neq \alpha_j \text{ pre } i \neq j, \text{ kde } i, j = 1, \ldots, 5.$$

Štatistika D_n patrí do kategórie U-štatistík popísaných v kapitole 2 (definícia 6). Jedna z vlastností D_n je symetria v $x_1, \ldots, x_n$. Funkcia $\phi(x_1, y_1; \ldots; x_5, y_5)$ je zároveň jadrom regulárneho funkcionálu Δ, takže štatistika D_n je nestranným odhadom Δ nad $\mathcal{F}$ a platí $\mathbb{E}D_n = \Delta$.

Použitím rovníc (2.2) a (3.5) môžeme štatistiku D_n vyjadriť nasledovne:

$$D_n = \binom{n}{5}^{-1} \sum_{R_{5,n}} D_5, \tag{3.6}$$

kde súčet prechádza cez α_i také, že $1 \leq \alpha_1 < \cdots < \alpha_5 \leq n$ a D_5 je symetrické jadro funkcionálu,

$$D_5 = \frac{1}{5!} \sum_{P_{5,n}} \phi(X_{\alpha_1}, Y_{\alpha_1}; \ldots; X_{\alpha_5}, Y_{\alpha_5}).$$

3.3.2 Výpočet štatistiky D

Použitím (3.4) a (3.5) môžeme štatistiku D_n ekvivalentne vyjadriť

$$D_n = \frac{A - 2(n-2)B + (n-2)(n-3)C}{n(n-1)(n-2)(n-3)(n-4)}, \tag{3.7}$$

kde

$$\begin{aligned} A &= \sum_{\alpha=1}^{n} a_\alpha(a_\alpha - 1)b_\alpha(b_\alpha - 1), \\ B &= \sum_{\alpha=1}^{n} (a_\alpha - 1)(b_\alpha - 1)c_\alpha, \\ C &= \sum_{\alpha=1}^{n} c_\alpha(c_\alpha - 1), \end{aligned} \tag{3.8}$$

a

$$\begin{aligned} a_\alpha &= \sum_{\beta=1}^{n} C(X_\alpha - X_\beta) - 1, \qquad b_\alpha = \sum_{\beta=1}^{n} C(Y_\alpha - Y_\beta) - 1, \\ c_\alpha &= \sum_{\beta=1}^{n} C(X_\alpha - X_\beta)C(Y_\alpha - Y_\beta) - 1, \end{aligned}$$

kde C je funkcia z (3.4).

Prvky $a_\alpha + 1$ a $b_\alpha + 1$ značia teda poradie X_α a Y_α; a c_α nám indikuje počet dvojrozmerných pozorovaní (X_β, Y_β), pre ktoré platí $X_\beta < X_\alpha$ a $Y_\beta < Y_\alpha$. V praxi sa na výpočet štatistiky D_n používa formula (3.7), pretože stačí vypočítať príslušné $a_\alpha, b_\alpha, c_\alpha$ pre každý člen náhodného výberu, dopočítať A, B, C z (3.8) a nakoniec vložiť do rovnice (3.7).

3.3.3 Odvodenie rozptylu

Zaujíma nás rozptyl testovej štatistiky D. Využijeme vyjadrenie z (3.6) a na začiatok sa zameriame na rozptyl D_5. Vďaka (2.2) vieme, že sa jedná o symetrické jadro funkcionálu.

Označme

$$\Phi(x_1, y_1; \ldots; x_5, y_5) = D_5 = \frac{1}{5!} \sum_{P_{5,n}} \phi(x_{\alpha_1}, y_{\alpha_1}; \ldots; x_{\alpha_5}, y_{\alpha_5}). \tag{3.9}$$

Predpokladajme, že rozptyl jadra $\Phi(x_1, y_1; \ldots; x_5, y_5)$ je konečný, tzn.

$$\operatorname{var}(\Phi(x_1, y_1; \ldots; x_5, y_5)) < \infty.$$

Ďalej zadefinujeme postupnosť funkcií Φ_k pridružených k jadru Φ.

Nech $k = 1, \ldots, 5$, označme

$$\begin{aligned} &\Phi_k(x_1, y_1; \ldots; x_k, y_k) = \mathbb{E}[\Phi(x_1, y_1; \ldots; x_k, y_k; X_{k+1}, Y_{k+1}; \ldots; X_5, Y_5)] = \\ &= \int \cdots \int \phi(x_1, y_1; \ldots; x_k, y_k; x_{k+1}, y_{k+1}; \ldots; x_5, y_5) dF(x_{k+1}, y_{k+1}) \cdots dF(x_5, y_5). \end{aligned} \tag{3.10}$$

Funkciu $\Phi_k(x_1, y_1; \ldots; x_k, y_k)$ môžeme chápať ako podmienenú strednú hodnotu $\Phi(x_1, y_1; \ldots; x_5, y_5)$ pri daných $(X_1, Y_1), \ldots, (X_k, Y_k)$.

Je zrejmé, že $\Phi_0 = \mathbb{E}[\Phi(X_1, Y_1; \ldots; X_5, Y_5)]$ a špeciálne v našom prípade pre $k = 5$ platí $\Phi_5(x_1, y_1; \ldots; x_5, y_5) = \mathbb{E}[\Phi(x_1, y_1; \ldots; x_5, y_5] = \Phi(x_1, y_1; \ldots; x_5, y_5)$.

Veta 10. *Funkcie Φ_k definované ako v (3.10) majú nasledujúce vlastnosti:*

1. $\Phi_k(x_1, y_1; \ldots; x_k, y_k) = \mathbb{E}[\Phi_d(x_1, y_1; \ldots; x_k, y_k; X_{k+1}, Y_{k+1}; \ldots; X_d, Y_d)]$ *pre* $1 \leq k < d \leq 5$,
2. $\mathbb{E}[\Phi_k(X_1, Y_1; \ldots; X_k, Y_k)] = \mathbb{E}[\Phi(X_1, Y_1; \ldots; X_5, Y_5)]$.

Dôkaz. Postupujeme ako v dôkaze Theorem 1 v Lee (1990).

Prvú rovnosť môžeme vyjadriť:

$$\begin{aligned}
&\mathbb{E}[\Phi_d(x_1, y_1; \ldots; x_k, y_k; X_{k+1}, Y_{k+1}; \ldots; X_d, Y_d)] = \\
&= \int \cdots \int \Phi_d(x_1, y_1; \ldots; x_k, y_k; x_{k+1}, y_{k+1}; \ldots; x_d, y_d) \prod_{i=k+1}^{d} dF(x_i) = \\
&= \int \cdots \int \{\int \cdots \int \Phi(x_1, y_1; \ldots; x_d, y_d; \ldots; x_5, y_5) \prod_{i=d+1}^{5} dF(x_i)\} \prod_{i=k+1}^{d} dF(x_i) = \\
&= \int \cdots \int \Phi(x_1, y_1; \ldots; x_5, y_5) \prod_{i=k+1}^{5} dF(x_i) = \\
&= \Phi_k(x_1, y_1; \ldots; x_k, y_k).
\end{aligned}$$

V druhom prípade platí

$$\begin{aligned}
&\mathbb{E}[\Phi_k(X_1, Y_1; \ldots; X_k, Y_k)] = \\
&= \int \cdots \int \Phi_k(x_1, y_1; \ldots; x_k, y_k) \prod_{i=1}^{k} dF(x_i) = \\
&= \int \cdots \int \{\int \cdots \int \Phi(x_1, y_1; \ldots; x_5, y_5) \prod_{i=k+1}^{5} dF(x_i)\} \prod_{i=1}^{k} dF(x_i) = \\
&= \int \cdots \int \Phi(x_1, y_1; \ldots; x_5, y_5) \prod_{i=1}^{5} dF(x_i) = \\
&= \mathbb{E}[\Phi(X_1, Y_1; \ldots; X_5, Y_5)].
\end{aligned}$$

❑

Špeciálne, z vety 10 máme

$$\Phi_{k-1}(x_1, y_1; \ldots; x_{k-1}, y_{k-1}) = \mathbb{E}[\Phi_k(x_1, y_1; \ldots; x_{k-1}, y_{k-1}; X_k, Y_k)].$$

Použitím lemmatu 7 a (2.2) môžeme vyjadriť strednú hodnotu D_5

$$\mathbb{E}[D_5] = \mathbb{E}[\Phi(X_1, Y_1; \ldots; X_5, Y_5)] = \Delta$$

a z vety 10 dostávame

$$\mathbb{E}[\Phi_k(X_1, Y_1; \ldots; X_k, Y_k)] = \Delta, \ k = 1, \ldots, 5. \tag{3.11}$$

Ďalej, pre $k = 1, \ldots, 5$ označíme rozptyl $\Phi_k(X_1, Y_1; \ldots; X_k, Y_k)$ ako

$$\sigma_k^2 = \mathrm{var}(\Phi_k(X_1, Y_1; \ldots; X_k, Y_k)), \quad \sigma_0^2 = 0,$$

čo môžeme vyjadriť aj nasledovne

$$\begin{aligned}\sigma_k^2 &= \mathbb{E}[\Phi_k(X_1, Y_1; \ldots; X_k, Y_k) - \Delta]^2 = \\ &= \mathbb{E}[\Phi_k(X_1, Y_1; \ldots; X_k, Y_k)]^2 - 2\Delta\mathbb{E}[\Phi_k(X_1, Y_1; \ldots; X_k, Y_k)] + \mathbb{E}[\Delta]^2 = \\ &= \mathbb{E}[\Phi_k(X_1, Y_1; \ldots; X_k, Y_k)]^2 - \Delta^2,\end{aligned}$$

kde sme využili (3.11).

Definícia 12. *Ak $\sigma_1^2 = 0$, povieme, že D_n je degenerovaná štatistika a príslušné jadro je degenerované jadro. V opačnom prípade sa jedná o nedegenerovanú štatistiku D_n a jej nedegenerované jadro.*

Rozptyl σ_k^2 podmienenej strednej hodnoty Φ_k môžeme interpretovať ako kovarianciu.

Veta 11. *Nech $\Phi(S) := \Phi(X_{i_1}, Y_{i_1}; \ldots; X_{i_5}, Y_{i_5})$, kde $S = \{i_1, \ldots, i_5\}$. Potom rozptyl σ_k^2 môžeme alternatívne vyjadriť*

$$\sigma_k^2 = \mathrm{cov}(\Phi(S_1), \Phi(S_2)),$$

kde S_1, S_2 sú dve 5-prvkové podmnožiny $\{1, \ldots, n\}$ s práve k spoločnými prvkami.

Dôkaz. Lee (1990, Theorem 2). ❑

Rozptyl D_n je daný nasledujúcim vzťahom:

$$\mathrm{var}(D_n) = \mathrm{var}\left(\binom{n}{5}^{-1} \sum_{\{i_1,\ldots,i_5\} \in R_{5,n}} \Phi(X_{i_1}, Y_{i_1}; \ldots; X_{i_5}, Y_{i_5})\right),$$

kde využívame (2.1), (3.6) a (3.9).

Užitočné vyjadrenie rozptylu štatistiky D môže byť skonštruované pomocou σ_k^2, čo bude ukázané v nasledujúcej vete.

Veta 12. *Nech D_n je U-štatistika s jadrom Φ. Potom*

$$\mathrm{var}(D_n) = \binom{n}{5}^{-1} \sum_{k=1}^{5} \binom{5}{k}\binom{n-5}{5-k}\sigma_k^2. \tag{3.12}$$

Dôkaz. Postupujeme ako v dôkaze Theorem 3 v Lee (1990).

$$\begin{aligned}&\mathrm{var}(D_n) = \mathrm{var}\left(\binom{n}{5}^{-1} \sum_{\{i_1,\ldots,i_5\} \in R_{5,n}} \Phi(X_{i_1}, Y_{i_1}; \ldots; X_{i_5}, Y_{i_5})\right) = \\ &= \binom{n}{5}^{-2} \mathrm{cov}\left(\sum_{R_{5,n}} \Phi(X_{i_1}, Y_{i_1}; \ldots; X_{i_5}, Y_{i_5})), \sum_{R_{5,n}} \Phi(X_{j_1}, Y_{j_1}; \ldots; X_{j_5}, Y_{j_5}))\right) = \\ &= \binom{n}{5}^{-2} \sum_{R_{5,n}} \sum_{R_{5,n}} \mathrm{cov}(\Phi(X_{i_1}, Y_{i_1}; \ldots; X_{i_5}, Y_{i_5}), \Phi(X_{j_1}, Y_{j_1}; \ldots; X_{j_5}, Y_{j_5})),\end{aligned}$$

kde suma prechádza cez všetky dvojice $\{i_1, \ldots, i_5\}$ a $\{j_1, \ldots, j_5\}$ 5-prvkových podmnožín množiny $\{1, 2, \ldots, n\}$.

Potrebujeme zistiť, koľko z týchto $\binom{n}{5}^2$ párov majú jeden prvok spoločný, dva prvky spoločné, atď. Potom budeme môcť použiť Vetu 11.

Uvažujme, koľkými spôsobmi môžeme vybrať dvojice 5-prvkových podmnožín, ktoré majú spoločných práve k prvkov. Očividne, prvého z dvojice môžeme vybrať ľubovoľne, takže máme $\binom{n}{5}$ možností. Ďalej vieme, že v dvojiciach má byť k spoločných prvkov, preto týchto k prvkov v druhej množine môžeme vybrať $\binom{5}{k}$ spôsobmi. Zvyšných $5-k$ prvkov vyberáme z $n-k-(5-k)$, kde $n-k$ sú možnosti, ktoré nám teoreticky ostali v množine $\{1, 2, \ldots, n\}$, ale $5-k$ sa má líšiť, aby mali skupiny spoločných len k prvkov. Preto počet dvojíc, ktoré majú k prvkov spoločných, a teda platí

$$\operatorname{cov}(\Phi(X_{i_1}, Y_{i_1}; \ldots; X_{i_5}, Y_{i_5}), \Phi(X_{j_1}, Y_{j_1}; \ldots; X_{j_5}, Y_{j_5})) = \sigma_k^2$$

je presne $\binom{n}{5}\binom{5}{k}\binom{n-5}{5-k}$. Potom

$$\begin{aligned}
&\binom{n}{5}^{-2} \sum_{R_{5,n}} \sum_{R_{5,n}} \operatorname{cov}(\Phi(X_{i_1}, Y_{i_1}; \ldots; X_{i_5}, Y_{i_5}), \Phi(X_{j_1}, Y_{j_1}; \ldots; X_{j_5}, Y_{j_5})) = \\
&= \binom{n}{5}^{-1} \sum_{k=1}^{5} \binom{5}{k}\binom{n-5}{5-k}\sigma_k^2.
\end{aligned}$$

Rovnosť (3.12) je dokázaná.

❑

Keďže D_n patrí do skupiny U-štatistík, ktoré bližšie rozobral Hoeffding, môžeme použiť už dokázané vlastnosti, ktoré platia pre rozptyl. Dôkazy k nasledujúcim vetám môžeme nájsť v Hoeffding (1948b).

Veta 13. *Funkcia* $n \operatorname{var}(D_n)$ *je klesajúca v* n *a*

$$\lim_{n\to\infty} n \operatorname{var}(D_n) = 25\sigma_1^2. \tag{3.13}$$

Veta 14. *Pre rozptyl* $\operatorname{var}(D_n)$ *platia nasledujúce nerovnosti*

$$\frac{25}{n}\sigma_1^2 \leq \operatorname{var}(D_n) \leq \frac{5}{n}\sigma_5^2.$$

Rozptyl testovej štatistiky D_n v prípade nezávislosti sa pomocou transformácie popísanej v Hoeffding (1948a) môže vyjadriť

$$\operatorname{var}(30D_n) = \frac{2(n^2+5n-32)}{9n(n-1)(n-3)(n-4)}. \tag{3.14}$$

Veta 15. *Nech* $\Delta = \mathbb{E}[\phi(X_1, Y_1; \ldots; X_5, Y_5)]$ *a* $\mathbb{E}[\phi(X_1, Y_1; \ldots; X_5, Y_5)]^2$ *sú konečné. Potom náhodná veličina* $\sqrt{n}(D_n - \Delta)$ *má asymptoticky normálne rozdelenie s nulovou strednou hodnotou a rozptylom* $25\sigma_1^2$.

Za predpokladu nezávislosti Hoeffding (1948a) ukázal, že $\sigma_1^2 = 0$, takže testová štatistika $\sqrt{n}D_n$ má limitne degenerované normálne rozdelenie, $\mathcal{N}(0, 0)$.

3.3.4 D-test nezávislosti

Skonštruujeme test nezávislosti. Nech α je požadovaná hladina významnosti a ρ_n je najmenšie číslo, ktoré splňuje nerovnosť

$$P(D_n > \rho_n | F \in \omega_0) \leq \alpha.$$

Vypočítame testovú štatistiku D_n pomocou (3.7). Nulovú hypotézu budeme zamietať práve vtedy, keď $D_n > \rho_n$. Pre $n = 5, 6, 7$ sú hodnoty ρ_n spočítané v Hoeffding (1948a). Ďalej v článku Mudholkar a Wilding (2003) môžeme nájsť kritické hodnoty za nulovej hypotézy pre konečné rozsahy výberov do $n = 100$.

Použitím Čebyševovej nerovnosti a (3.14) dostaneme

$$P\left(30D_n > \sqrt{\frac{2(n^2+5n-32)}{9n(n-1)(n-3)(n-4)\alpha}}\right) \leq \alpha,$$

takže platí

$$30\rho_n \leq \sqrt{\frac{2(n^2+5n-32)}{9n(n-1)(n-3)(n-4)\alpha}}.$$

Teda ρ_n konverguje k nule pro $n \to \infty$.

Preto ak $\Delta > 0$ (platí H_1), potom $\Delta - \rho_n > 0$ pre dostatočne veľké n a platí

$$P(D_n > \rho_n) \geq P(|D_n - \Delta| \leq \Delta - \rho_n) \geq 1 - \frac{\operatorname{var} D_n}{(\Delta - \rho_n)^2}. \tag{3.15}$$

Dá sa nahliadnuť, že použitím (3.13), konverguje pravá strana k 1, a teda sila testu konverguje k 1. Z (3.15) a z vety 9 dostávame, že D-test je konzistentný vzhľadom k triede $\mathcal{F}$.

4. Simulácie

Priebeh simulácie pozostáva z nasledujúcich krokov:

1. Na začiatku zvolíme dané $\alpha \in (0, 1)$, ktoré bude značiť predpokladanú hladinu testu.

2. Nasimulujeme 1000-krát náhodný výber o rozsahu $n \in \{8, 16, 24, 32, 48, 64\}$.

Bude nás zaujímať, pre aké rozsahy výberu je odhadnutá hladina testu dostatočne blízka predpokladanej. Taktiež sa pozrieme na prípady, kedy je hladina porušená a o aké veľké porušenie sa jedná.

Simulácie predvedieme v štatistickom software R Core Team (2016). Tento software spolu s Microsoft Excel využijeme na grafické znázornenie výsledkov.

Porovnávať budeme 5 testov, ktoré sme popísali v predchádzajúcich kapitolách. Jedná sa o Pearsonov korelačný test, Spearmanov korelačný test, Kendallov korelačný test, test založený na korelácii vzdialenosti a Hoeffdingov D-test. Z prostredia R využijeme pri výpočtoch knižnice Hmisc (implementácia Hoeffdingovho testu), energy (na výpočet korelácie vzdialenosti) a AUC (na výpočet plochy pod ROC krivkou, ako bude vysvetlené nižšie). Na výpočet p-hodnôt sme používali funkciu cov.test príslušnej metódy – Pearson, Spearman, Kendall; dcov.test pre test založený na korelácii vzdialenosti a hoeffd pre Hoeffdingovu metódu.

4.1 Postup

Simulačná štúdia pozostáva z dvoch častí. V prvej časti sa zrealizuje 1000 simulácii pre každý z nasledujúcich prípadov:

1. Nezávislé náhodné veličiny: náhodné veličiny X_i a Y_i sú generované nezávisle na sebe z normálneho rozdelenia, $X_i \sim \mathcal{N}(0, 2)$, $Y_i \sim \mathcal{N}(0, 1)$ pre $i = 1, 2, \ldots, n$ a rozsahy $n \in \{8, 16, 24, 32, 48, 64\}$.

2. Lineárna transformácia: X_i, $i = 1, 2, \ldots, n$, je náhodný výber z normálneho rozdelenia ako v bode 1. a Y_i je lineárnou funkciou X_i. Náhodnú veličinu Y_i môžeme zapísať $Y_i = f(X_i) + \epsilon_i$, kde f je lineárna funkcia a $\epsilon_i \sim \mathcal{N}(0, 0.5)$ značí chybu merania. Konkrétne, v našej simulácii pracujeme s $Y_i = 0.5X_i + \epsilon_i$. Rozsahy lineárnej transformácie a všetkých ďalších transformácií sú $n \in \{8, 16, 24, 32, 48, 64\}$.

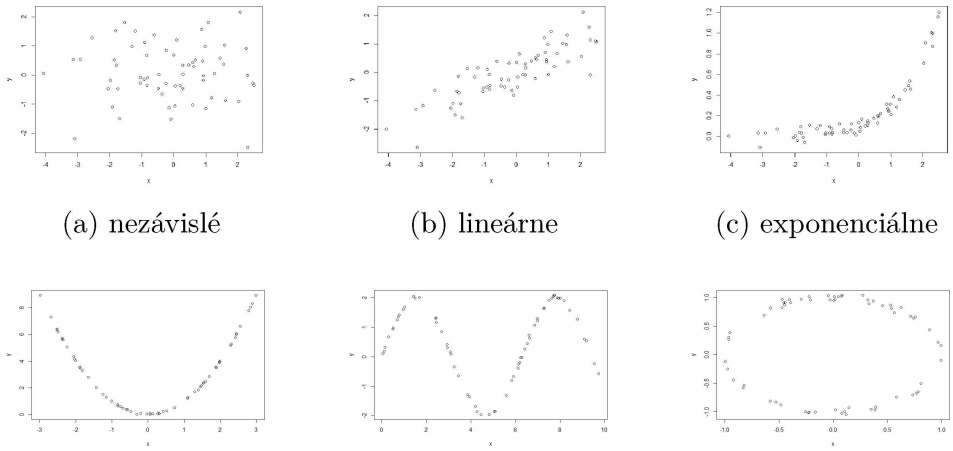
3. Exponenciálna transformácia: X_i, $i = 1, 2, \ldots, n$, je náhodný výber z normálneho rozdelenia ako v bode 1. a Y_i je exponenciálnou funkciou X_i. Náhodnú veličinu Y_i môžeme zapísať $Y_i = f(X_i) + \epsilon_i$, kde f je exponenciálna funkcia a $\epsilon_i \sim \mathcal{N}(0, 0.5)$ značí chybu merania. V simulácii pracujeme s $Y_i = 0.1 \exp(X_i) + \epsilon_i$.

4. Kvadratická transformácia: X_i, $i = 1, 2, \ldots, n$, je náhodný výber z rovnomerného rozdelenia, $X_i \sim \mathcal{R}(-3, 3)$ a Y_i je kvadratickou funkciou X_i. Náhodnú veličinu Y_i môžeme zapísať $Y_i = f(X_i) + \epsilon_i$, kde f je kvadratická funkcia a $\epsilon_i \sim \mathcal{N}(0, 0.5)$ značí chybu merania. V simulácii používame $Y_i = X_i^2 + \epsilon_i$.

5. Transformácia sínus: X_i, $i = 1, 2, \ldots, n$, je náhodný výber z rovnomerného rozdelenia, $X_i \sim \mathcal{R}(0, 10)$ a Y_i je v tvare $Y_i = 2\sin(X_i) + \epsilon_i$, kde $\epsilon_i \sim \mathcal{N}(0, 0.5)$ značí chybu merania.

6. Transformácia kružnica: X_i, $i = 1, 2, \ldots, n$, je náhodný výber z rovnomerného rozdelenia, $X_i \sim \mathcal{R}(-1, 1)$ a Y_i je v tvare $Y_i = \delta_i\sqrt{1 - X_i^2} + \epsilon_i$, kde $\delta_i \sim \mathcal{R}(\{-1, 1\})$ má rovnomerné rozdelenie na dvojprvkovej množine $\{-1, 1\}$ a $\epsilon_i \sim \mathcal{N}(0, 0.5)$ značí chybu merania.

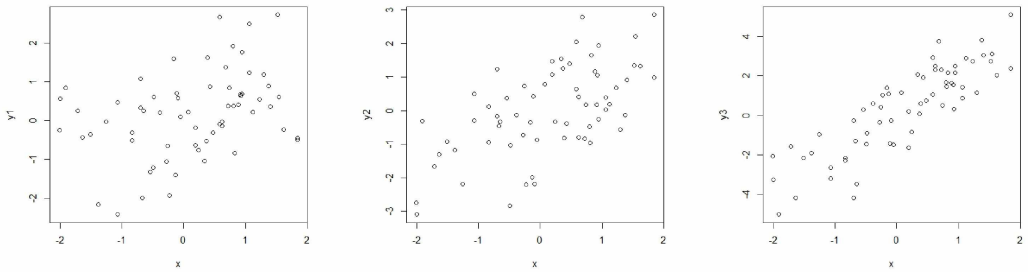
Transformácie boli vybrané tak, aby sme zahrnuli lineárne monotónne prípady, ktoré sú reprezentované lineárnou transformáciou, nelineárne monotónne prípady reprezentované exponenciálnou transformáciou, nelineárne nemonotónne prípady reprezentované kvadratickou transformáciou a použitím funkcie sinus a nakoniec pomocou transformácie kružnicou. Jednotlivé vzťahy medzi náhodnými veličinami X_i a Y_i sú vyzobrazené na obrázku 4.1.



(a) nezávislé (b) lineárne (c) exponenciálne

(d) kvadratické (e) sínus (f) kružnica

Obrázek 4.1: Simulácia 1. Vzťahy medzi veličinami X a Y. Ukážky realizácií náhodných výberov X_i (vodorovná os) a Y_i (zvislá os), $i = 1, \ldots, 64$, pre 6 skúmaných prípadov.



(a) $\mathrm{corr}(X, Y) = 0.35$ (b) $\mathrm{corr}(X, Y) = 0.61$ (c) $\mathrm{corr}(X, Y) = 0.86$

Obrázek 4.2: Simulácia 2. Lineárna závislosť medzi veličinami X a Y podľa veľkosti korelácie.

V druhej časti budeme pracovať s lineárnou transformáciu, pričom sa pozrieme, ako sa mení empirická sila testov pri zmene korelácie medzi X a Y. Náhodná veličina X_i pochádza z normovaného normálneho rozdelenia, $X_i \sim \mathcal{N}(0,1)$, a v tomto prípade pracujeme s chybou pri meraní $\epsilon_i \sim \mathcal{N}(0,1)$. Obecne, zvolíme Y_i v tvare $Y_i = aX_i + \epsilon_i$. Potom $\text{var}(Y_i) = \text{var}(aX_i + \epsilon_i) = \text{var}(aX_i) + \text{var}(\epsilon_i) = a^2\text{var}(X_i) + \text{var}(\epsilon_i) = a^2 + 1$ a

$$\text{corr}(X_i, Y_i) = \frac{\text{cov}(X_i, Y_i)}{\sqrt{\text{var}(X_i)}\sqrt{\text{var}(Y_i)}} = \frac{a\text{cov}(X_i, X_i)}{\sqrt{a^2+1}} = \frac{a}{\sqrt{a^2+1}},$$

kde sme využili vlastnosti korelačného koeficientu opísané vo vete 4 a nezávislosť X_i a ϵ_i.

Rozlíšime tri prípady podľa veľkosti korelačného koeficientu medzi X a Y. Volíme rôzne $a_j \in \mathbb{R}$, $j = 1, 2, 3$, kde $a_1 = 0.2$, $a_2 = 0.8$, $a_3 = 2$. Jednotlivé vzťahy sú zobrazené na obrázku 4.2.

4.2 Výsledky

Cieľom bolo zistiť, ktoré zo spomínaných testov dokážu rozoznať závislosť či nezávislosť medzi náhodnými veličinami X a Y. V tabuľke 4.1 môžeme nájsť výsledky simulácií prevedené na náhodné výbery o rozsahu $n \in \{8, 16, 24, 32, 48, 64\}$ za nulovej hypotézy, tj. v prípade nezávislosti X a Y. Pre každú metódu a každé n sme spočítali empirickú hladinu testu.

n	Pearson	Spearman	Kendall	Dcor	Hoeffding
8	0.054	0.063	0.075	0.056	0.131
16	0.040	0.044	0.049	0.044	0.089
24	0.045	0.053	0.047	0.047	0.061
32	0.046	0.045	0.043	0.043	0.053
48	0.055	0.057	0.053	0.060	0.064
64	0.045	0.056	0.053	0.047	0.057

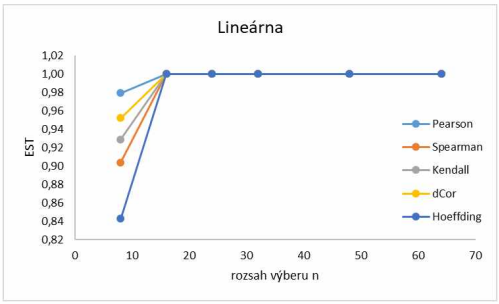
Tabulka 4.1: Empirická hladina testu.

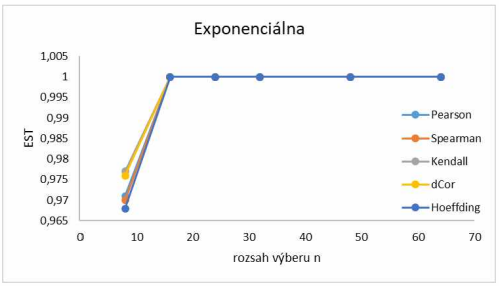
V prípade závislosti boli využité lineárne aj nelineárne transformácie a zamerali sme sa na jednotlivé sily testov v prípade transformácií 2 až 6. Brali sa do úvahy rozsahy výberov $n \in \{8, 16, 24, 32, 48, 64\}$ a pre každý z nich, pre každú metódu a pre každé porušenie nezávislosti bola vypočítaná empirická sila daných testov (viď obrázok 4.3 a 4.4).

Ďalej sme sa zamerali na priemernú p-hodnotu testov pri porušení nezávislosti (viď tabuľka 4.2). V prípade, že je porušená nezávislosť (neplatí nulová hypotéza), chceli by sme čo najviackrát zamietnuť H_0. Z teórie testovania hypotéz, rozoberanej napr. v Anděl (2007) máme, že H_0 zamietame práve vtedy, keď je p-hodnota menšia než predom stanovená hladina α. Toto pravidlo využijeme pri posudzovaní, ktorý z testov si viedol pri testovaní nezávislosti lepšie.

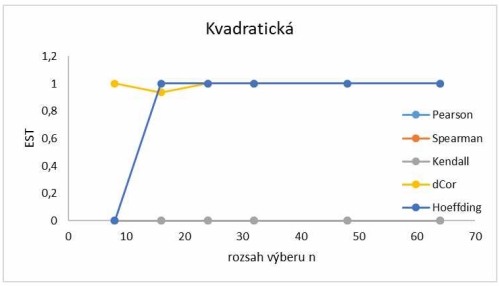
Na záver tejto časti bola zhotovená ROC krivka[1] a nasledovne vypočítaná plocha pod touto krivkou, nazývaná AUC[2] (viď tabuľka 4.3). Jedná sa o ďalšiu

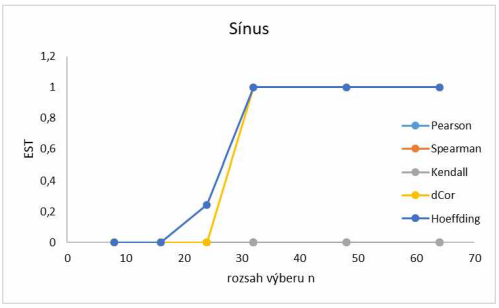
[1] z anglického receiver operating characteristic curve
[2] z anglického area under ROC curve





Obrázek 4.3: Empirická sila testu podľa typu závislosti dvoch náhodných veličín (lineárna, exponenciálna) a použitej metódy (Pearson, Spearman, Kendall, Dcor, Hoeffding).





Obrázek 4.4: Empirická sila testu podľa typu závislosti dvoch náhodných veličín (kvadratická, sinus) a použitej metódy (Pearson, Spearman, Kendall, Dcor, Hoeffding).

pomôcku pri vyhodnocovaní sily testov. Obecne, ide o dvojdimenzionálny graf, ktorý ukazuje vzťah medzi senzitivitou a špecificitou testu, pričom na x-ovej osy je vyzobrazená $1-$špecificita a y-ová os predstavuje senzitivitu. Senzitivita testu, inak nazývaná aj citlivosť testu, vyjadruje úspešnosť, s ktorou test zachytil prítomnosť sledovaného stavu u daného subjektu. Na druhú stranu, špecificita vyjadruje schopnosť testu správne vybrať prípady, u ktorých skúmaný znak nenastáva.

V našom prípade sú na x-ovej osy hodnoty $\alpha \in (0, 1)$ a na y-ovej osy je zobrazená $h(\alpha) = \frac{1}{1000} \sum_{i=1}^{1000} \mathbf{1}\{p\text{-}hodnota_i < \alpha\}$, $h(\alpha) \in [0, 1]$, inak povedané pomer prípadov, keď zamietame H_0 a všetkých prípadov, ktoré nastali. Funkcia $h(\alpha)$ predstavuje ROC krivku a plochu pod touto krivkou vypočítame použitím Riemannovho súčtu. Platí, že plocha pod ROC krivkou, AUC, dosahuje hodnoty z intervalu $[0, 1]$. Ak je hodnota AUC blízka 1, môžeme povedať, že sila testu je vysoká. Plocha pod krivkou rovná 0.5 značí, že test dáva výsledky ekvivalentné hodu mincou. V prípade, že je AUC menšia ako 0.5, testovaná metóda nie je schopná rozlíšiť závislosť medzi náhodnými veličinami. Výsledky našich simulácií sú rozdelené podľa typu závislosti medzi náhodnými veličinami X a Y, podľa veľkosti náhodného výberu a použitej metódy (viď tabuľka 4.3).

V druhej časti simulácií sme vypočítali empirickú silu testu v závislosti od veľkosti hodnoty korelačného koeficientu pre jednotlivé metódy a rozsahy výberov $n \in \{8, 16, 24, 32, 48, 64\}$ (viď tabuľka 4.4).

vzťah	n	Pearson	Spearman	Kendall	Dcor	Hoeffding
	8	$6.63 \cdot 10^{-3}$	$1.79 \cdot 10^{-2}$	$2.16 \cdot 10^{-2}$	$1.18 \cdot 10^{-2}$	$5.24 \cdot 10^{-2}$
lin	16	$7.19 \cdot 10^{-5}$	$2.67 \cdot 10^{-4}$	$7.78 \cdot 10^{-4}$	$1.27 \cdot 10^{-3}$	$3.44 \cdot 10^{-4}$
	32	$5.78 \cdot 10^{-11}$	$1.92 \cdot 10^{-10}$	$1.83 \cdot 10^{-8}$	$9.99 \cdot 10^{-4}$	$1.00 \cdot 10^{-8}$
	64	$9.23 \cdot 10^{-14}$	$5.20 \cdot 10^{-12}$	$4.06 \cdot 10^{-11}$	$9.99 \cdot 10^{-4}$	$1.00 \cdot 10^{-8}$
	8	$1.43 \cdot 10^{-2}$	$7.76 \cdot 10^{-3}$	$1.07 \cdot 10^{-2}$	$1.31 \cdot 10^{-2}$	$1.09 \cdot 10^{-2}$
exp	16	$1.35 \cdot 10^{-4}$	$1.93 \cdot 10^{-6}$	$1.07 \cdot 10^{-5}$	$9.99 \cdot 10^{-4}$	$3.84 \cdot 10^{-8}$
	32	$1.55 \cdot 10^{-7}$	$1.51 \cdot 10^{-12}$	$1.29 \cdot 10^{-10}$	$9.99 \cdot 10^{-4}$	$1.00 \cdot 10^{-8}$
	64	$3.43 \cdot 10^{-14}$	$3.85 \cdot 10^{-23}$	$2.05 \cdot 10^{-19}$	$9.99 \cdot 10^{-4}$	$1.00 \cdot 10^{-8}$
	8	$1.69 \cdot 10^{-1}$	$5.40 \cdot 10^{-1}$	$6.65 \cdot 10^{-1}$	$2.65 \cdot 10^{-2}$	$1.05 \cdot 10^{-1}$
quad	16	$8.27 \cdot 10^{-1}$	$9.36 \cdot 10^{-1}$	$8.69 \cdot 10^{-1}$	$4.04 \cdot 10^{-2}$	$2.29 \cdot 10^{-3}$
	32	$5.10 \cdot 10^{-1}$	$8.58 \cdot 10^{-1}$	$8.15 \cdot 10^{-1}$	$8.83 \cdot 10^{-3}$	$7.55 \cdot 10^{-7}$
	64	$5.59 \cdot 10^{-1}$	$5.71 \cdot 10^{-1}$	$5.09 \cdot 10^{-1}$	$1.05 \cdot 10^{-3}$	$1.00 \cdot 10^{-8}$
	8	$9.46 \cdot 10^{-1}$	$8.74 \cdot 10^{-1}$	$8.37 \cdot 10^{-1}$	$4.77 \cdot 10^{-1}$	1.00
sine	16	$5.15 \cdot 10^{-1}$	$2.83 \cdot 10^{-1}$	$2.60 \cdot 10^{-1}$	$2.63 \cdot 10^{-1}$	$1.15 \cdot 10^{-1}$
	32	$4.31 \cdot 10^{-1}$	$3.73 \cdot 10^{-1}$	$4.29 \cdot 10^{-1}$	$1.21 \cdot 10^{-2}$	$3.35 \cdot 10^{-4}$
	64	$7.47 \cdot 10^{-1}$	$2.72 \cdot 10^{-1}$	$1.78 \cdot 10^{-1}$	$4.98 \cdot 10^{-3}$	$1.91 \cdot 10^{-6}$
	8	$5.13 \cdot 10^{-1}$	$4.51 \cdot 10^{-1}$	$5.42 \cdot 10^{-1}$	$7.48 \cdot 10^{-1}$	$1.95 \cdot 10^{-1}$
circ	16	$4.43 \cdot 10^{-1}$	$6.69 \cdot 10^{-1}$	$8.56 \cdot 10^{-1}$	$2.76 \cdot 10^{-1}$	$1.71 \cdot 10^{-1}$
	32	$4.10 \cdot 10^{-1}$	$6.19 \cdot 10^{-1}$	$8.18 \cdot 10^{-1}$	$3.40 \cdot 10^{-1}$	$7.94 \cdot 10^{-2}$
	64	$9.26 \cdot 10^{-1}$	$7.49 \cdot 10^{-1}$	$7.52 \cdot 10^{-1}$	$2.16 \cdot 10^{-1}$	$4.17 \cdot 10^{-3}$

Tabulka 4.2: Priemerná p-hodnota podľa rozsahu výberu a typu závislosti dvoch náhodných veličín. Označenie závislosti medzi X a Y: lin – lineárna, exp – exponenciálna, quad – kvadratická, sine – sínus, circ – kružnica.

4.3 Diskusia

V prípade nezávislých náhodných veličín sme pracovali s náhodným výberom o rozsahu $n \in \{8, 16, 24, 32, 48, 64\}$. Ako hlavný indikátor budeme považovať empirickú hladinu testu (tabuľka 4.1). Analyzovaním počtu zamietnutí H_0 relatívne k celkovému počtu pozorovaní dostávame, že sa všetky použité metódy blížia k predpísanej hladine $\alpha = 0.05$. Môžeme si všimnúť, že Hoeffdingov test pre menšie rozsahy výberov dosahuje približne dvojnásobne vyššie hodnoty než stanovená hladina, ale so zvyšujúcim sa n sa tieto hodnoty približujú k α.

Pozrieme sa na prípady, keď existuje nejaká závislosť medzi náhodnými veličinami. Pri lineárnej transformácii sa najlepšie preukázal Pearsonov test, ktorý dosiahol najvyššie hodnoty empirickej sily a najnižšie priemerné p-hodnoty, ale obecne všetky testy dokazujú veľkú silu, ktorá vychádza 1 už pri rozsahu $n = 16$ rovnako ako plocha pod ROC krivkou. Pri exponenciálnej (nelineárnej, monotónnej) transformácii najnižšie priemerné p-hodnoty preukázal Spearmanov a Kendallov test (tabuľka 4.2). Opakovane však môžeme povedať, že testy dosahujú vysokej sily a hodnoty AUC blízke 1 pri relatívne malých rozsahoch výberu (viď obrázok 4.3, tabuľka 4.3). Môžeme to pripísať tomu, že v oboch prípadoch sa jedná o monotónnu transformáciu medzi náhodnými veličinami a všetky spomínané testy sú schopné ju odhaliť.

Pri nemonotónnych transformáciách, ako je napríklad kvadratická, sinus či kružnica, Pearsonov, Spearmanov a Kendallov test neboli schopné odhaliť vzťah

vzťah	n	Pearson	Spearman	Kendall	Dcor	Hoeffding
	8	0.99	0.98	0.98	0.99	0.95
lin	16	1.00	1.00	1.00	1.00	1.00
	32	1.00	1.00	1.00	1.00	1.00
	64	1.00	1.00	1.00	1.00	1.00
	8	0.99	0.99	0.99	0.99	0.99
exp	16	1.00	1.00	1.00	1.00	1.00
	32	1.00	1.00	1.00	1.00	1.00
	64	1.00	1.00	1.00	1.00	1.00
	8	0.83	0.46	0.34	0.97	0.90
quad	16	0.17	0.06	0.13	0.96	1.00
	32	0.49	0.14	0.19	0.99	1.00
	64	0.44	0.43	0.49	1.00	1.00
	8	0.05	0.13	0.16	0.52	0.00
sine	16	0.48	0.72	0.74	0.74	0.88
	32	0.57	0.63	0.57	0.99	1.00
	64	0.25	0.73	0.82	1.00	1.00
	8	0.49	0.55	0.46	0.25	0.80
circ	16	0.56	0.33	0.14	0.72	0.83
	32	0.59	0.38	0.18	0.66	0.92
	64	0.07	0.25	0.25	0.78	1.00

Tabulka 4.3: Plocha pod krivkou (AUC) podľa rozsahu výberu a typu závislosti dvoch náhodných veličín.

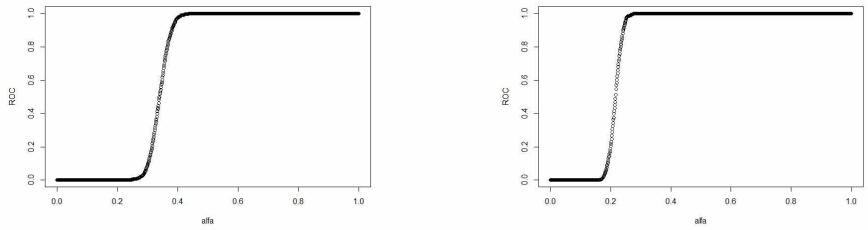
medzi veličinami nezávisle od rozsahu výberu. Na druhú stranu, pre kvadratickú transformáciu a transformáciu sínus Hoeffdingov test a test založený na korelácii vzdialenosti dokázali identifikovať určitú závislosť medzi náhodnými veličinami, pričom Hoeffdingov test si v našich simuláciách viedol o niečo lepšie (viď tabuľka 4.2 a 4.3). Tieto dva testy sa preukázali byť konzistentné vzhľadom k rozsahu výberu nielen pre lineárne či monotónne transformácie, ale taktiež pre nemonotónne (viď obrázok 4.4). Taktiež priemerné p-hodnoty pre výbery $n = 16$ a vyššie pre obe metódy v našich simuláciách vychádzajú menšie než predpísaná hladina α. Ďalej si môžeme všimnúť, že čím väčší rozsah výberu n, tým väčší vychádza obsah pod ROC krivkou a rovnako máme väčšiu empirickú sila testu. Zvyšné testy ukazujú hodnoty zväčša menšie či blízke ako 0.5, čo značí, že nie sú schopné rozoznať nemonotónnu závislosť medzi veličinami.

Pre nefunkčnú transformáciu, ktorá je zastúpená transformáciou pomocou kružnice, si viedol najlepšie Hoeffdingov test, ktorý ako jediný dokázal odhaliť takúto formu závislosti a pre ktorý vychádzajú priemerné p-hodnoty menšie než predpísaná hladina. Treba však podotknúť, že aj použitím tejto metódy je potrebný o niečo väčší výber. Analyzovaním AUC sa utvrdzujeme, že Hoeffdingov test najlepšie zvláda identifikovať nefunkčné transformácie. Pre test založený na korelácii vzdialenosti hodnoty AUC tiež stúpajú a približujú sa 1, môžeme preto predpokladať, že pre väčšie výbery by táto metóda bola schopná vo väčšine prípadov úspešne odhaliť takúto formu závislosti (obrázok 4.5).

Pri analyzovaní výsledkov lineárnych transformácií, pri ktorých sme sledovali rôzne hodnoty korelácie medzi veličinami X a Y, bolo ukázané, tak ako sme

corr	n	Pearson	Spearman	Kendall	Dcor	Hoeffding
	8	0.06	0.07	0.08	0.07	0.13
0.35	16	0.14	0.13	0.14	0.12	0.16
	32	0.23	0.21	0.21	0.21	0.21
	64	0.36	0.34	0.34	0.31	0.32
	8	0.38	0.33	0.37	0.33	0.40
0.61	16	0.93	0.89	0.88	0.89	0.87
	32	1.00	1.00	1.00	0.99	0.99
	64	1.00	1.00	1.00	1.00	1.00
	8	0.98	0.91	0.93	0.96	0.85
0.86	16	1.00	1.00	1.00	1.00	1.00
	32	1.00	1.00	1.00	1.00	1.00
	64	1.00	1.00	1.00	1.00	1.00

Tabulka 4.4: Empirická sila testu podľa rozsahu výberu a hodnoty korelácie medzi dvomi náhodnými veličinami.

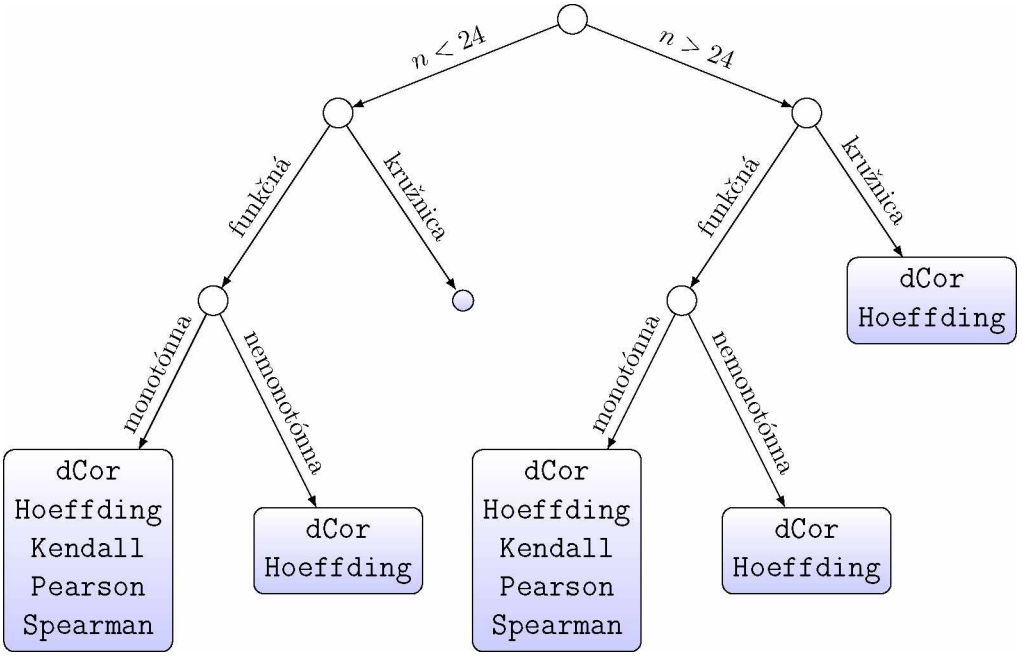


(a) ROC pre $n = 32$ (b) ROC pre $n = 64$

Obrázek 4.5: ROC krivka pre metódu dCor a transformáciu kružnicou.

očakávali, že čím väčšia hodnota korelačného koeficientu, tým bola empirická sila jednotlivých metód vyššia. V prípade Pearsonovho testu a testu založenom na korelácii vzdialenosti pre corr = 0.86 dosiahla empirická sila najvyššie hodnoty aj pre menšie výbery, pričom všetky metódy dosiahli veľmi dobré výsledky. Najlepšie si viedol Pearsonov test, ktorý vo väčšine prípade (pre vyššie hodnoty korelácie) dosahoval najväčšej sily zo všetkých spomínaných testov. Tento výsledok môžeme vysvetliť tak, že Pearsonov test sa špecializuje na identifikovanie lineárnych transformácií, kdežto ostatné testy sú obecnejšie a pokrývajú aj iné typy závislosti medzi veličinami.

Výsledky simulácií sú zhrnuté na obrázku 4.6, kde je ukázané, aký typ závislosti sú jednotlivé metódy schopné rozoznať vzhľadom k rozsahu náhodného výberu n a podľa toho, či sa jedná o funkčné monotónne alebo nemonotónne transformácie či nefunkčné transformácie reprezentované kružnicou.



Obrázek 4.6: Rozhodovací strom. Prehľad, aké vzťahy medzi náhodnými veličinami X a Y dokážu jednotlivé metódy identifikovať v závislosti od rozsahu výberu a typu závislosti medzi veličinami. Metódy sú zoradené abecedne.

5. Záver

Cieľom bakalárskej práce bolo prezentovať problém testovania nezávislosti dvoch náhodných veličín v neparametrickom modeli spojitých distribučných funkcií. Na začiatku boli predstavené základné pojmy a tvrdenia z teórie nezávislosti a z oblasti testovania založenom na poradí. Ďalej sme opísali neparametrické metódy používané na testovanie nezávislosti, konkrétne Spearmonov test, Kendallov test, test založený na korelácii vzdialenosti a Hoeffdingov test, ktoré sme spoločne s Pearsonovým testom medzi sebou porovnali. Zaujímal nás hlavne výsledok Hoeffdingovho neparametrického testu nezávislosti, ktorý bol v práci podrobnejšie opísaný. Skúmaním empirickej hladiny testu sme zistili, že všetky testy za nulovej hypotézy, teda v prípade nezávislosti náhodných veličín, dosiahli hodnoty blízke predpísanej hladiny α.

V prípade, že bola nezávislosť porušená, sústredili sme sa na to, ktoré metódy sú schopné odhaliť dané porušenia. Vybrali sme 5 typov porušenia nezávislosti, ktoré sme analyzovali:

1. Prípad, keď existuje lineárna závislosť medzi X a Y,
2. Prípad, keď existuje exponenciálna závislosť medzi X a Y,
3. Prípad, keď existuje kvadratická závislosť medzi X a Y,
4. Prípad, keď vzťah medzi X a Y môžeme opísať funkciou sínus,
5. Prípad, keď vzťah medzi X a Y môžeme opísať kružnicou.

Zistili sme, že monotónne transformácie (lineárnu a exponenciálnu) sú schopné rozoznať všetky spomenuté testy. Zmena nastáva pri nemonotónnych transformáciách, ako napríklad kvadratická transformácia či sinus, kde sa osvedčila metóda založená na korelácii vzdialenosti a Hoeffdingova metóda. Zvyšné tri metódy neboli schopné identifikovať takýto typ závislosti, rovnako ako neboli schopné rozoznať transformáciu kružnicou. Takúto formu závislosti medzi náhodnými veličinami sa v našich simuláciách pre väčšie rozsahy výberu podarilo odhaliť Hoeffdingovej metóde, pričom môžeme predpokladať, že pri dostatočne veľkých rozsahoch výberu metóda založená na korelácii vzdialenosti by bola schopná taktiež identifikovať takýto vzťah.

Na záver sme skúmali, ako si vedú testy pri zmene veľkosti korelačného koeficientu. Pozreli sme sa na tri prípady: keď je medzi náhodnými veličinami veľmi malá lineárna závislosť, stredná a keď náhodné veličinu prejavujú väčšie hodnoty lineárnej závislosti. Obecne, všetky testy dokázali odhaliť lineárnu závislosť pri vyšších hodnotách korelačného koeficientu. Pearsonov test preukázal najlepšie výsledky aj pre menšie rozsahy výberu, čo môžeme pripísať tomu, že testy, ktoré sú konzistentné vzhľadom k väčšej triede alternatív, budú mať menšiu silu vzhľadom k podtriede alternatív (ako napríklad lineárna) než testy, ktoré majú optimálne vlastnosti vzhľadom k danej podtriede.

V ďalšom výskume by bolo zaujímavé sledovať, ako sa mení sila jednotlivých testov pri nemonotónnych a nefunkčných transformáciách pre väčšie rozsahy výberov, špeciálne u metódy založenej na korelácii vzdialenosti a Hoeffdingovej metódy, pričom výsledky by nám mohli poskytnúť lepšiu predstavu o jednotlivých metódach.

Literatura

Anděl, J. (2007). *Základy matematické statistiky.* Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

Blum, J. R., Kiefer, J. a Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.*, **32**, 485–498. doi: 10.1214/aoms/1177705055.

Halmos, P. R. (1946). The theory of unbiased estimation. *Ann. Math. Statist.*, **17**, 34–43. doi: 10.1214/aoms/1177731020.

Hoeffding, W. (1948a). A non-parametric test of independence. *Ann. Math. Statist.*, **19**, 546–557. doi: 10.1214/aoms/1177730150.

Hoeffding, W. (1948b). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, **19**, 293–325. doi: 10.1214/aoms/1177730196.

Hotelling, H. a Pabst, M. R. (1936). Rank correlation and tests of significance involving no assumption of normality. *Ann. Math. Statist.*, **7**, 29–43. doi: 10.1214/aoms/1177732543.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, **30**, 81–93. doi: 10.2307/2332226.

Lee, A. J. (1990). *U-statistics : Theory and Practice.* Statistics: A series of Textbooks and Monographs. M. Dekker, M., New York. ISBN 9780824782535.

Mudholkar, G. S. a Wilding, G. E. (2003). On the conventional wisdom regarding two consistent tests of bivariate independence. *J. Roy. Statist. Soc. Ser. D (The Statistician)*, **52**, 41–57. doi: doi:10.1111/1467-9884.00340.

Omelka, M. (2018). NMSA331 Matematická statistika 1, Poznámky k přednášce. URL https://www.karlin.mff.cuni.cz/~omelka/Soubory/nmsa331/ms1_1617.pdf.

R Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Székely, G. J., Rizzo, M. L. a Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.*, **35**, 2769–2794. doi: 10.1214/009053607000000505.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge. ISBN 9780521784504.

Seznam obrázků

4.1 Simulácia 1. Vzťahy medzi veličinami X a Y. Ukážky realizácií náhodných výberov X_i (vodorovná os) a Y_i (zvislá os), $i = 1, \ldots, 64$, pre 6 skúmaných prípadov. 20
4.2 Simulácia 2. Lineárna závislosť medzi veličinami X a Y podľa veľkosti korelácie. 20
4.3 Empirická sila testu podľa typu závislosti dvoch náhodných veličín (lineárna, exponenciálna) a použitej metódy (Pearson, Spearman, Kendall, Dcor, Hoeffding). 22
4.4 Empirická sila testu podľa typu závislosti dvoch náhodných veličín (kvadratická, sinus) a použitej metódy (Pearson, Spearman, Kendall, Dcor, Hoeffding). 22
4.5 ROC krivka pre metódu dCor a transformáciu kružnicou. 25
4.6 Rozhodovací strom. Prehľad, aké vzťahy medzi náhodnými veličinami X a Y dokážu jednotlivé metódy identifikovať v závislosti od rozsahu výberu a typu závislosti medzi veličinami. Metódy sú zoradené abecedne. 26

Seznam tabulek

4.1 Empirická hladina testu. 21
4.2 Priemerná p-hodnota podľa rozsahu výberu a typu závislosti dvoch náhodných veličín. Označenie závislosti medzi X a Y: lin – lineárna, exp – exponenciálna, quad – kvadratická, sine – sínus, circ – kružnica. 23
4.3 Plocha pod krivkou (AUC) podľa rozsahu výberu a typu závislosti dvoch náhodných veličín. 24
4.4 Empirická sila testu podľa rozsahu výberu a hodnoty korelácie medzi dvomi náhodnými veličinami. 25